

Preprocessing, Variable Selection, and Classification Rules in the
Application of SIMCA Pattern Recognition to Mass Spectral Data

Illinois Univ. at Chicago Circle

Prepared for:

Environmental Protection Agency, Research Triangle Park, NC

1989



TECHNICAL REPORT DATA (Please read instructions on the reverse before completing)		
1. REPORT NO EPA/600/J-89/316	2.	3. RECI PB90-198722
4. TITLE AND SUBTITLE Preprocessing, Variable Selection and Classification Rules in the Application of SIMCA Pattern Recognition to Mass Spectral Data	5. REPORT DATE	
7. AUTHOR(S) W. J. Dunn III, S. L. Emery and W. Graham Glen University of Illinois at Chicago, D. R. Scott-EPA	6. PERFORMING ORGANIZATION CODE	
9. PERFORMING ORGANIZATION NAME AND ADDRESS U. S. Environmental Protection Agency Atmospheric Research and Exposure Assessment Laboratory Methods Research and Development Division Research Triangle Park, N. C. 27711	8. PERFORMING ORGANIZATION REPORT NO.	
12. SPONSORING AGENCY NAME AND ADDRESS U. S. Environmental Protection Agency Office of Research and Development 401 M. Street, S.W. Washington, D.C. 20460	10. PROGRAM ELEMENT NO.	11. CONTRACT/GRANT NO. CR 814363
15. SUPPLEMENTARY NOTES To be submitted to Environmental Science and Technology	13. TYPE OF REPORT AND PERIOD COVERED Journal Article	
16. ABSTRACT In a recent report a strategy was proposed for the classification and identification of toxic organic compounds observed in ambient air from mass spectra using computational pattern recognition based on SIMCA principal components modeling of the autocorrelation transformed mass spectra. With this technique very good classification and identification results (87% and 84%, respectively) were obtained with GC/MS from training and calibration data for the 78 toxic compounds targeted for routine monitoring in ambient air. However, when applied to GC/MS ambient air field data, a number of hydrocarbons were incorrectly classified as chlorocarbons indicating that the training sets were not optimal for discriminating between these classes. A new strategy for data preprocessing, variable selection and class model optimization has been developed to solve this problem. Only the sixteen most intense ions in each mass spectrum are retained. The MS data are scaled by taking the square root of the intensities and the autocorrelation transform is then taken. A training class has been introduced for hydrocarbons in addition to three other classes. The original SIMCA classification rule has been modified to give a more reasonable approximation of the training set pattern structure and object distances from the class models. The results of these changes gave a similar classification accuracy for the 122 training spectra (98%) but greatly improved results for real data (94-97%).	14. SPONSORING AGENCY CODE EPA/600/09	
17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field Group
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED 20. SECURITY CLASS (This page) UNCLASSIFIED	21. NO. OF PAGES 9 22. PRICE A02

Reprinted from ENVIRONMENTAL SCIENCE & TECHNOLOGY, Vol. 23, Page 1499, December 1989
Copyright © 1989 by the American Chemical Society and reprinted by permission of the copyright owner.

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

Preprocessing, Variable Selection, and Classification Rules in the Application of SIMCA Pattern Recognition to Mass Spectral Data

William J. Dunn, III,* Silvio L. Emery,[†] and W. Graham Glen

College of Pharmacy, The University of Illinois at Chicago, 833 South Wood, Chicago, Illinois 60612

Donald R. Scott

Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency,
Research Triangle Park, North Carolina 27711

■ In a recent report a strategy was proposed for the classification and identification of toxic organic compounds observed in ambient air from mass spectra using computational pattern recognition based on SIMCA principal-components modeling of the autocorrelation-transformed mass spectra. With this technique very good classification and identification results (87 and 84%, respectively) were obtained with GC/MS from training and calibration data for the 78 toxic compounds targeted for routine monitoring in ambient air. However, when applied to GC/MS ambient air field data, a number of hydrocarbons were incorrectly classified as chlorocarbons, indicating that the training sets were not optimal for discriminating between these classes. A new strategy for data preprocessing, variable selection, and class model optimization has been developed to solve this problem. Only the 16 most intense ions in each mass spectrum are retained. The MS data are scaled by taking the square root of the intensities and the autocorrelation transform is then taken. A training class has been introduced for hydrocarbons in addition to other classes. The original SIMCA classification rule has been modified to give a more reasonable approximation of the training set pattern structure and object distances from the class models. The results of these changes gave a similar classification accuracy for the 122 training spectra (98%) but greatly improved results for real data (94–97%).

Introduction

A large number of environmental problems require analytical methods based on gas chromatography/mass spectrometry (GC/MS). The U.S. Environmental Protection Agency, for example, uses computer-supported gas chromatography/quadrupole mass spectrometry as its primary monitoring tool (1). In most environmental ap-

plications the analytical method of choice, GC/MS, will lead to complex spectral data due to the many components observed in samples and the multidimensional nature of the analysis. While the mixtures are complex, the number of target compounds sought in such mixtures may only be in the range of 100–150. The target list approach makes the interpretation less difficult, and it can be greatly improved with an appropriate computer-assisted scheme. To the present time, however, the emphasis has been placed mainly on analytical methods development with very little emphasis being placed on software development for data interpretation. The result is that the analyst must rely on available commercial software resident on the monitoring instrument. This software has been designed for general applications and is not well-suited, in most cases, for atypical analytical problems, which are the rule in environmental monitoring situations, e.g., complex mixtures in which the components occur at very low concentrations. In a recent review of the analysis of hazardous organic compounds by GC/MS, it was suggested that today's instruments and computer programs are inadequate for interpretation of the resulting data (2).

Environmental surveillance and monitoring is data acquisition and interpretation intensive. The interpretation of such data is not a trivial problem and quite different from single spectra identification. Large volumes of data are generated that require considerable computer time, even on a minicomputer, for interpretation. Therefore, it would be an advantage to be able to perform the data interpretation on a computer other than being used for instrument control and data acquisition. Recently, mass spectral data bases with software for identification of unknown mass spectra have become available for use on microcomputers (3, 4), including the NBS/EPA/NIH data base and the Wiley registry of mass spectral data. These identification methods use library searching techniques that require large libraries of mass spectra and are rela-

* Present address: Chemical Research and Development Command, Aberdeen Proving Grounds, MD 21040.

tively slow in actual practice. Special mass storage equipment is required, and only limited information can be extracted from the mass spectrum of an unknown if it is not in the data base. Also, the identification accuracy on complex environmental samples with some of the commercial software can be low (5).

In this laboratory we have been developing software for application to complex environmental GC/MS data (6, 7). For such complex data these classification and identification techniques have a number of advantages over commercial software currently available. The programs can be run on microcomputers and are computationally based. Since they utilize target sets of compounds the approaches do not require a large data base, as library searching techniques do.

The approach is based on SIMCA pattern recognition (8). SIMCA, Simple Modeling by Class Analogy, is a class modeling method that provides a chemical class assignment for an unknown. This may be sufficient for some environmental applications in cases where identification is difficult (9). We have recently applied this technique to the mass spectra of a target list of 78 compounds sought in ambient air. The pattern recognition method was trained on spectra from the CIS/EPA/NIH mass spectral library from an INCOS system and applied to calibration and field data obtained during the process of ambient air monitoring (6, 7). Very good class assignment (87%) and identification (84%) results were obtained with the training and calibration data (6). For those target compounds in the field data that were included in the training sets, classification and identification accuracies of 88 and 85% were obtained (7). For nontarget compounds in the field samples, which were primarily hydrocarbons, a number of false positives were obtained. Because of this misclassification problem, the techniques have been reexamined and improved. In this report an approach to data preprocessing, variable selection, and class model optimization is presented, which gives excellent results when applied to mass spectral data from two actual field samples. Only the class assignment aspects of the data analysis are addressed in the paper.

Methods

All computations were carried out on an IBM PC/XT microcomputer equipped with an 8087 math coprocessor and a 20 + 20 megabyte Bernoulli box. A version of SIMCA pattern recognition, UNiversity of Illinois PARTial Least Squares, or UNIPALS, was written in BASIC. A description of the software is available from the authors. Library, calibration, and field data were downloaded from an INCOS library or from Finnigan GC/MS files and stored on the PC/XT. The data were reformatted and submitted to analysis by using the approach previously discussed (6, 7). The 78 target and other training compounds are listed in Table I.

Four classes of compounds are represented in the target and training sets (6). Class 1 is the nonhalogenated benzenes, class 2 is the chlorine-containing compounds, class 3 is the bromo and bromochloro compounds, and class 4 is the aliphatic hydrocarbons. We also include in class 4 miscellaneous oxygen-containing hydrocarbon-like compounds. These include a few aliphatic alcohols, aldehydes, and ketones with spectra very similar to their aliphatic hydrocarbon counterparts. Such compounds are frequently present in ambient air. The highest mass in the training set was 256 with an interval from 35 to 256 m/z .

Classification and Identification

Our previous reports discussed (6, 7) the scheme on

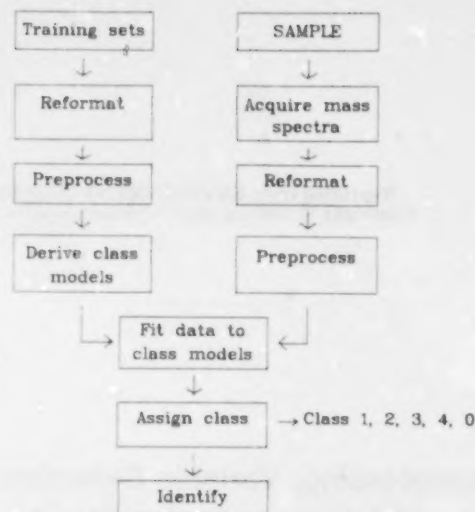


Figure 1. Scheme for application of pattern recognition to mass spectral data.

which this computer-assisted approach to mass spectral classification and compound identification was based. A detailed description will not be presented here, but the scheme is shown in Figure 1.

Pattern recognition studies are done in steps, as shown in the scheme. In the top left side of the scheme, training sets are selected and the data are reformatted. In the next step the data are preprocessed. This is a very important, but little discussed, aspect of pattern recognition. The actual type of preprocessing to be applied is a function of the form of the original data. We have chosen to first scale the mass spectral intensities, which are normalized to a base peak, by taking the square root of the intensities. This has the effect of giving less weight to the very few ions of high intensity in a spectrum and more weight to the more numerous ions of low intensity. Preprocessing in this way has the effect of scaling the data and is subject to the same risks.

The library spectra used for training had been clipped so that they generally contain only 16 peaks in each. A typical mass spectrum from a field sample contains approximately 40 nonzero ions. Therefore, only the 16 largest peaks were retained for each field spectrum to facilitate comparisons with the training data.

After the spectra were scaled and clipped, the data were further preprocessed to provide the autocorrelation transform (10) of the mass spectra. This transform has been suggested by McGill and Kowalski (11) and Wold and Christie (12). In these reports and our earlier ones (6, 7), the mean-corrected transform was used. Here, the spectral mean was not subtracted prior to applying the transform. The particular transform is given in eq 1.

$$r_j = \sum (I_i I_{i+j}) / \sum (I_i)^2 \quad (1)$$

I_i is the intensity of ion i , I_{i+j} is the intensity of ion $i + j$, and r_j is the autocorrelation coefficient for ions separated by j mass units. j is referred to as the lag at which the coefficient is calculated. In this study, the maximum lag was 100, which neglects neutral fragments with mass greater than 100 thereby assuming that these are not significant. In addition, the statistical significance of lags above 100 tends to diminish due to the small number of product terms, which appear in the numerator of the transform.

Table I. Target List and Training Compounds^a

compd no.	name	compd no.	name
1	benzene	62	2-bromo-1-propene
2	methylbenzene	63	3-bromo-1-propene
3	benzonitrile	64	1-bromopropane
4	ethenylbenzene	65	2-bromopropane
5	1,2-dimethylbenzene	66	1-bromobutane
6	1,3-dimethylbenzene	67	bromobenzene
7	1,4-dimethylbenzene	68	dibromomethane
8	ethylbenzene	69	1,2-dibromoethane
9	benzaldehyde	70	1,2-dibromopropane
10	(1-methylethyl)benzene	71	1,3-dibromopropane
11	1,3,5-trimethylbenzene	72	2,2-dibromopropane
12	1-phenethanone	73	1,3-dibromobutane
13	butylbenzene	74	1,4-dibromobutane
14	1-methyl-4-(1-methylethyl)benzene	75	tribromomethane
15	3-chloro-1-propene	76	bromoethane
16	1-chloropropane	77	2-methylpropane
17	2-chloropropane	78	butane
18	1-chlorobutane	79	2-methylbutane
19	2-chlorobutane	80	pentane
20	1-chloro-2,3-epoxypropane	81	4-methyl-1-pentene
21	2-chloroethoxyethene	82	2,3-dimethylbutane
22	chlorobenzene	83	2-methylpentane
23	1-chloro-2-methylbenzene	84	hexane
24	1-chloro-3-methylbenzene	85	2,3,3-trimethyl-1-butene
25	1-chloro-4-methylbenzene	86	(E)-2-heptene
26	1-ethenyl-4-chlorobenzene	87	3-ethylpentane
27	1,1-dichloroethene	88	heptane
28	1,1-dichloroethane	89	4,5-dimethyl-1-hexene
29	1,2-dichloroethane	90	3-methyleneheptene
30	1,2-dichloropropane	91	(Z)-2-octene
31	1,3-dichloropropane	92	2,2,3,3-tetramethylbutane
32	3,4-dichloro-1-butene	93	2,2,3-trimethylpentane
33	cis-1,4-dichloro-2-butene	94	2,3-dimethylhexane
34	1,3-dichlorobutane	95	2,4-dimethylhexane
35	1,4-dichlorobutane	96	3,4-dimethylhexane
36	2,2-dichlorobutane	97	3-methylheptane
37	2,3-dichlorobutane	98	4-methylheptane
38	1,2-dichlorobenzene	99	octane
39	1,3-dichlorobenzene	100	2,2-dimethylheptane
40	1,4-dichlorobenzene	101	2,3-dimethylheptane
41	trichloromethane	102	2,5-dimethylheptane
42	trichloroethene	103	4-ethylheptane
43	1,1,1-trichloroethane	104	nonane
44	1,1,2-trichloroethane	105	cyclohexane
45	1,1,2-trichloropropane	106	methylcyclohexane
46	1,2,3-trichloropropane	107	cis-1,2-dimethylcyclohexane
47	tetrachloromethane	108	ethylcyclohexane
48	tetrachloroethene	109	1,2,4-trimethylcyclopentane (1a,2a,4b)
49	1,1,1,2-tetrachloroethane	110	1,2,3-trimethylcyclopentane (1a,2a,4b)
50	1,1,2,2-tetrachloroethane	111	1,1,3,3-tetramethylcyclopentane
51	1,1,1,2-tetrachloropropane	112	2-methylpropylcyclopentane
52	1,1,2,3-tetrachloropropane	113	cis-1-ethyl-2-methylcyclohexane
53	1,2,2,3-tetrachloropropane	114	trans-1-ethyl-2-methylcyclohexane
54	pentachloroethane	115	1,1,3-trimethylcyclohexane
55	bromochloromethane	116	1,3,5-trimethylcyclohexane
56	1-bromo-1-chloroethane	117	1,3,5-trimethylcyclohexane (1a,3a,5a)
57	1-bromo-2-chloroethane	118	1,2,5-trimethylcyclohexane (1a,3a,5b)
58	1-bromo-3-chloropropane	119	cis-1-ethyl-4-methylcyclohexane
59	bromodichloromethane	120	tetrahydrofuran ^b
60	bromotrichloromethane	121	1,4-dioxane ^b
61	dibromochloromethane		

^aCompounds 1-76 are target compounds. ^bNot used in training.

By use of the transformed spectra, principal-components models were derived for the four training sets with eq 2.

$$r_{kj} = r_j + \sum_{a=1}^A t_{ka} b_{aj} + e_{kj} \quad (2)$$

Here, k is a compound index, j is the mass interval index, and A is the number of principal components or product terms in the model. r_j is the mean for the autocorrelation coefficient for ions separated by j mass units. The e_{kj} are

the model residuals. From the residuals a class model standard deviation is calculated and from this an interval around each model can be defined in which there is a high probability of finding a member of that class. This is the basis for a class assignment with the approach.

In order to find an optimal subset of variables for class modeling, modeling power, MPOW, was calculated for each variable (6). Thirty-eight variables that had MPOW of <0.10 in all models were deleted. After deletion of the irrelevant variables, new models were derived by using the

Table II. Classification Results of Field Data

scan no.	library search	visual inspection	visual	SIMCA	revised SIMCA
262	NO ₂	Field Sample 1 (590)			
266	?	CO ₂ + NO ₂	0	0	0
269	2,2-dimethyloxetane	CO ₂	0	4	0
277	2,2,4,4-tetramethyl-1,3-cyclobutadione	2-methylpropane + CO ₂	4	4	4
281	5-methyl-2-hexanone	3-methyl-1-butene + CO ₂	4	0	0
285	?	butane	4	4	4
287	pentane	2-propanol + 2-pentene	4	0	0
291	?	pentane	4	4	4
294	1,1,2-trichloro-1,2,2-trifluoroethane	dichloromethane	2	2	2
300	?	1,1,2-trichloro-1,2,2-trifluoroethane	2	0	2
303	2-methyl-1-propen-1-one	?	?	4	4
305	methoxytrimethylsilane	2-methyl-1-propen-1-one	4	4	4
306	2,3-dimethylbutane	methoxytrimethylsilane	0	4	4
312	ethoxyethene	2-methylpentane	4	4	4
315	2,3,3-trimethylhexane	2-butanone	4	4	4
323	3-methylpentane	3-methylpentane	4	4	4
334	?	hexane	4	4	4
338	perfluorotoluene	?	?	4	4
344	1,1,1-trichloroethane	perfluorotoluene	0	1	0
353	1-aminocyclopentanecarboxylic acid	1,1,1-trichloroethane	2	2	2
355	benzene	?	?	4	0
359	?	benzene	1	1	1
362	2-methyl-1-pentene	carbon tetrachloride + hydrocarbon	2, 4	0	1
370	O-(2-methylpropyl)hydroxylamine	cyclohexane and/or 2-methyl-1-pentene	4	4	4
376	?	2,3-dimethylpentane	4	4	4
386	?	hydrocarbon + dibromo compound	4, 3	4	4
390	2,2,3,3-tetramethylbutane	5-methylhexene + dichloro compound	4, 2	2	4
401	heptane	2,2,3,3-tetramethylbutane	4	4	4
403	?	heptane	4	4	4
406	2,4-dimethylcyclobutane	3-ethyl-1-pentene	4	4	4
411	3-ethyl-2-pentene	(E)-2-heptene	4	4	4
415	1-heptene	3-heptene	4	4	4
419	cyclopentanecarboxaldehyde	2-heptene	4	4	4
425	methylcyclohexane	(E)-2-hexenal	4	4	4
439	3-ethyl-4-methylhexane	methylcyclohexane	4	4	4
442	2,4-dimethylhexane	?	?	0	0
451	5-methyl-1-heptene	2,2-dimethylpentane	4	1	4
465	?	3-methyleneheptane	4	4	4
471	toluene	?	?	4	4
481	2,3-dimethylhexane	toluene	1	2	1
490	2-methylheptane	2,3-dimethylhexane	4	4	4
494	4-methyl-2-pentanone	2-methylheptane	4	4	4
500	?	4-methyl-2-pentanone	4	4	4
502	?	?	?	4	4
505	cis-3,4-dimethylcyclohexane	1,1,1-trichloropropane + hydrocarbon	2, 4	4	4
508	hexanal	3-heptene-2-one	4	4	4
521	2,2,5-trimethylhexane	hexanal	4	4	4
524	3-methyleneheptane	2,2-dimethylheptane	4	4	4
527	1-octene	3-methyleneheptane	4	4	4
538	trans-1,2-dimethylcyclohexane	1-octene	4	4	4
545	tetrachloroethene	cis-1,2-dimethylcyclohexane	4	4	4
549	2,3,4-trimethylhexane	tetrachloroethene + hydrocarbon	2, 4	2	2
555	2-octene	octane	4	4	4
562	5-methyl-1-heptene	(Z)-2-octene	4	4	4
570	4-octene	3-methyleneheptane	4	4	4
595	1-fluorodecane	(Z)-4-octene	4	4	4
597	?	chloro compound + hydrocarbon	2, 4	4	4
601	hexamethylcyclotrisilane	chloro compound + hydrocarbon	2, 4	0	2
606	1-methyl-3-(1-methylethyl)cyclopentane	?	0	0	2
617	1,1,3-trimethylcyclohexane	ethylcyclohexane	4	4	4
622	2,3,6-trimethylheptane	1,1,3-trimethylcyclohexane	4	4	4
632	?	2,5-dimethylheptane	4	4	4
643	ethylbenzene	?	?	0	0
649	1,3,5-trimethylcyclohexane	ethylbenzene	1	1	1
652	?	1,3,5-trimethylcyclohexane + hydrocarbon	4	4	4
663	1,2-dimethylbenzene	?	?	0	0
687	3-heptanone	1,2-dimethylbenzene	1	1	1
701	bicyclo[4.2.0]octa-1,3,5-triene	3-heptanone	4	4	4
708	5-methyleneundecane	styrene + hydrocarbon	1, 4	2	0
712	1,2-dimethylbenzene	2,6-dimethyl-1-heptene	4	4	4
716	heptanal	1,2-dimethylbenzene	1	1	1
728	2-methyl-1-octene	heptanal + hydrocarbon	4	4	4
740	1-nonene	2-methyl-1-octene	4	4	4
753	4-methyl-3-heptene	1-nonene	4	4	4
		4-methyl-3-heptene	4	4	4

Table II (Continued)

scan no.	library search	visual inspection	visual	SIMCA	revised SIMCA
761	3,3,5-trimethyl-1-hexene	(2-methylpropyl)cyclopentane	4	4	4
769	4-methyloctane	nonane	4	4	4
776	1-nonene	4-nonene	4	4	4
780	?	?	3	3	3
790	isopropylbenzene	isopropylbenzene	1	1	1
795	4-nonene	4-nonene	4	4	4
830	propylcyclohexane	propylcyclohexane	4	4	4
926	3-ethyl-5-methylheptane	3-ethyl-5-methylheptane	4	4	4
863	propylbenzene	propylbenzene	1	1	1
878	?	?	?	2	4
883	1-ethyl-3-methylbenzene	1-ethyl-3-methylbenzene	1	1	1
888	1-ethyl-2-methylbenzene	1-ethyl-2-methylbenzene	1	1	1
904	1-ethyl-3-methylbenzene	1,3,5-trimethylbenzene	1	1	1
926	?	hydrocarbon mixture	1, 4	0	0
928	2,3,6-trimethylheptane	2,3,6-trimethylheptane + 1-ethyl-4-methylbenzene	1, 4	0	4
934	2-ethyl-1-decanol	2-methylnonane	4	4	4
950	2,3,7-trimethyloctane	2-methylnonane	4	4	4
960	?	?	?	4	4
966	1-ethyl-3-methylbenzene	1,2,4-trimethylbenzene	1	1	1
974	octanal	octanal	4	4	4
983	?	?	2	2	2
995	cyclopropane	1-decene	4	4	4
1026	decane	decane	4	4	4
1032	?	1,2,3-trimethylbenzene + hydrocarbon	1, 4	4	4
1034	1,3,5-trimethylbenzene	1,3,5-trimethylbenzene	1	0	0
1053	(E)-2-decene	(E)-2-decene	4	0	0
1082	1-fluoro-4-iodobenzene	1-fluoro-4-iodobenzene	0	2	2
1089	3-ethyl-octane	hydrocarbon + 1-fluoro-4-iodobenzene	4, 0	4	4

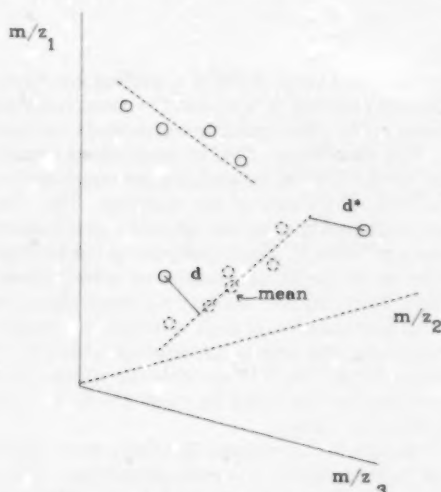


Figure 2. Geometric interpretation of SIMCA pattern recognition classification rule. d , distance of object from class model when object has a principal-component score within the range of the class model; d^* , distance of object from class model when the object has a principal-component score that is outside the range of the model.

subset of 62 variables and these were applied to both the training and field data.

One of the objectives of this approach was to obtain a class assignment for an unknown mass spectrum if it was not a member of the target list. A geometric representation of the usual SIMCA classification rule is illustrated in Figure 2. Here two classes and their models are shown. Since we are working at pattern recognition level two (13), three classification results are possible for an unknown. A mass spectral vector can be assigned to one of the two training sets, or it can be a member of neither.

With the SIMCA classification rule, class membership is determined by the orthogonal projection distance, d , of the unknown to the class models. In the case where the

unknown is beyond the class window determined by the extreme principal-component scores for the training data, the distance, d^* , is calculated from the unknown to the edge of the classes. The distances d and d^* , when corrected for differences in degrees of freedom, can be directly compared with the class residual standard deviation, s_{res} , as defined in eq 4. The e_{kj} are those given in eq 2. If the unknown is "similar" to the class of training compounds, d or d^* will be approximately equal to s_{res} . An approximate F statistic can be calculated to determine the level of significance of similarity and, therefore, of the classification result. This similarity rule has been discussed elsewhere (6), and a variation of it has been proposed (14).

The SIMCA classification rule gives equal weight to objects with projections near the extremes of the class and those near the geometrical mean or centroid of the class. Principal-component scores, the t 's in eq 2, are the positions of the objects in the models. In figure 2 the distances of two objects to the same class model are compared and under the usual classification rule they are equidistant from the model. SIMCA thus gives them the same classification result.

The t 's contain information about class assignment. In order to have the t 's considered in the classification rule, a variation of the usual SIMCA classification rule is proposed and used in this study.

Pattern or feature space can be divided into two subspaces (14). That subspace defined by the b vectors in eq 2 is the inside model space or IMS. The remaining axes we refer to as outside model space (OMS) axes or OMS axes. The root-mean-square variance along each b vector is given in eq 3. Here t_{ka} is the principal-component score for a compound in component a and the distance from the center of the class model to the point of projection of the object onto the class model. m is the number of compounds in the class. The standard deviation of the t 's along axis a is given by

$$s_a = [(1/(m - a)) \sum_k t_{ka}^2]^{1/2} \quad (3)$$

Table III. Summary of Classification Results

sample	classification rule			
	SIMCA		revised SIMCA	
	2.0 rsd	2.2 rsd	2.0 rsd	2.2 rsd
training data	121/122	120/122	119/122	119/122
field sample 1	72/80	73/80	75/80	75/80
field sample 2	135/150	138/150	146/150	146/150
total	328/352	331/352	340/352	340/352
total, %	93	94	97	97

Accuracy for EPA Target Compounds

sample	revised SIMCA rule	
	false positives (classes 1-3)	comps missed (classes 1-3)
training data	0/47	1/75
field sample 1	1/63	1/17
field sample 2	1/107	3/43
total	2/217	5/135

The remaining root-mean-square variance, or OMS distance, is calculated by eq 4. Here A is the number of

$$s_{\text{eps}} = [(1/(p-A))(1/(m-A-1))\sum_k \sum_j e_{kj}^2]^{1/2} \quad (4)$$

principal-component axes in the class model, p is the number of variable axes in feature space so that $(p-A)$ is the number of OMS axes, and t_{ka} is the principal-component score for compound k projected onto axis A . The summation is taken over both compounds and variables.

In the revised SIMCA classification study, class models are derived and unknowns are then fitted to the various class models. An unknown compound when fitted to an A -component model will have scores $t_{k1}, t_{k2}, \dots, t_{kA=A}$ and an OMS distance, d_{OMS} of

$$d_{\text{OMS}} = [(1/(p-A))\sum_k e_{kj}^2]^{1/2} \quad (5)$$

and an IMS distance, d_{IMS} , of

$$d_{\text{IMS}} = [(1/A)\sum_a t_{ka}^2 (s_{\text{eps}}/s_a)^2]^{1/2} \quad (6)$$

and a total distance

$$d = \frac{3}{4}d_{\text{OMS}} + \frac{1}{4}d_{\text{IMS}} \quad (7)$$

The OMS distance is calculated as in SIMCA. If the unknown actually belongs to the class whose model it is being fitted to, then $d_{\text{OMS}} \approx s_{\text{eps}}$. Otherwise, $d_{\text{OMS}} \gg s_{\text{eps}}$. The residuals are smallest when the unknown is fitted to its correct class. The IMS distance, d_{IMS} , is different from that calculated by SIMCA. With SIMCA, a class window is defined by the class model. If the projection of the unknown spectrum lies within this window, then its t_{ka} is set to zero. If its projection lies outside the class window its t_{ka} is equal to that at the edge of the window. In practice, the projections of nearly all of the unknowns lie within the SIMCA class windows, hence, the $d_{\text{IMS}} = 0$. Therefore, the d_{IMS} provides little or no class discrimination in SIMCA.

The modified d_{IMS} as given in eq 6 has been found to be useful in improving class discrimination, particularly by reducing the number of false positive classifications in class 2, the chlorocarbon class.

In the classification step, the new distance d_n for a compound is calculated to each of the class models. The unknown is then assigned to the nearest class, i.e., the one with the smallest d , provided that (i) $d < 2.0 s'_{\text{eps}}$, where s'_{eps} is the average residual standard deviation of the

training classes, and (ii) the smallest d is at least 8% smaller than any of the other d 's. If either condition is not met the unknown is assigned to class 0. The numerical factors in conditions i and ii were chosen on the basis that they give optimal classification results for these data and the new distance formula. For the original SIMCA classification distances, the optimal cutoff is found to be $d < 2.2 s'_{\text{eps}}$. This new classification rule is illustrated in Figure 3.

Results

The revised and usual SIMCA classification rules were applied to the training data and to the mass spectra from two field samples. The training compounds are listed in Table I. The ambient air field compounds as tentatively identified by the INCOS forward library search system are given in Table II for one of the samples. The class assignment as determined by our revised classification rule is also given in Table II. Each spectrum in the field sample was displayed on the PC and compared with its spectrum in the NBS/EPA/MSDC standard reference data base for the PC. An identification of each spectrum was made from this comparison and this is included in Table II. Only those unknowns in the field samples for which a definite visual class assignment could be made are used in scoring the classification results.

For classification purposes, the data were scaled by taking the square root of the mass intensities. The autocorrelation transform was applied to the 122 training spectra, and these data were then used to derive four SIMCA class models. In all cases models with two principal components were derived that explained 50-75% of the variance in the class data. These models were used to classify the training and unknown compound spectra by using the standard SIMCA classification rule and its variant as discussed above.

A summary of the classification results is given in Table III. For the training data, the usual SIMCA rule classified compounds accurately at a rate of 98% (120/122) by using the optimal cutoff distance of 2.2 standard deviations (SD). The revised rule was comparable with a correct classification result of 98% (119/122) with a cutoff distance 2.0 SD from the class models.

However, there was a significance difference in the classification accuracy of the two rules when applied to field data. In this case, the correct class was determined from a visual identification of the unknown spectra. Only those spectra that on inspection did not appear to be

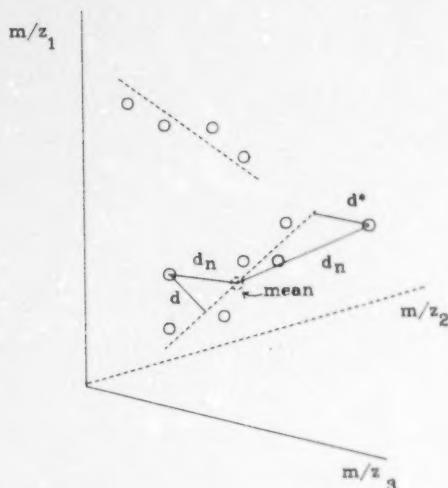


Figure 3. Geometric interpretation of the revised SIMCA rule. The symbols are the same as in Figure 2. d_n , new distance of the objects from Figure 2 to the class mean.

mixtures were considered in the analysis. These are noted in Table II.

The usual SIMCA rule was 91% (73/80) accurate when applied to field sample 1 and 92% (138/150) accurate when applied to field sample 2. The revised rule had a classification accuracy of 94% (74/80) on field sample 1 and 97% (146/150) when applied to field sample 2. Thus, the overall error rate is significantly lower with the revised SIMCA rule (9 errors in the 230 field spectra as compared to 19 errors with SIMCA).

Four modifications of the previous classification techniques (10) were made in the present study. The mass spectral data were scaled differently; the autocorrelation transform was varied; the classification rule was modified; and the interfering hydrocarbons were included in a separate class. Judging from the results of the training set data in Table III, the change in classification rule had little effect, comparing the results on the training data from our previous study (7) with those here. The field sample results are much better than those obtained previously. It is difficult to identify a single factor that led to the improvement. It is clear that very good results can be obtained with the present overall classification procedure, especially with real field data.

The library data on which the method was trained are ideal spectra and contain little noise, as indicated from the

fact that the training spectra all are with 1.5 SD of their models. The field data were obtained on different instruments and contain noise from interfering substances. The revised rule can tolerate more experimental variation in the data than the usual SIMCA rule.

Acknowledgments

We greatly acknowledge the assistance of Lynn Wright, Curtis Morris, and Joe Bumgarner of EMSL, U.S. EPA, Research Triangle Park, NC, by providing GC/MS data files. We thank the Research Resources Center of the University of Illinois at Chicago for the use of the data processing unit of the Finnigan 4510 for data transfer to the PC.

Literature Cited

- (1) Gurka, D. F.; Donnelly, B.; Hinnert, T. A.; Heithmar, E. M.; Titus, R.; Henshaw, J. M. *Anal. Chem.* 1988, 60, 454A.
- (2) Swallow, K. C.; Shifrin, N. S.; Doherty, P. J. *Environ. Sci. Technol.* 1988, 22, 136.
- (3) McLafferty, F. W.; Stauffer, D. B. *J. Chem. Inf. Comput. Sci.* 1985, 25, 245.
- (4) Martinsen, D. P. *Appl. Spectrosc.* 1981, 35, 255.
- (5) Pellizzari, E. D.; Hartwell, T.; Crowder, J. U.S. Environmental Protection Agency Report PB-85-125664. Research Triangle Park, NC, 1985.
- (6) Dunn, W. J., III; Koehler, M. G.; Emery, S. L.; Scott, D. R. *Chemom. Intell. Lab. Syst.* 1987, 1, 321.
- (7) Scott, D. R.; Dunn, W. J., III; Emery, S. *Environ. Sci. Technol.* 1987, 21, 891.
- (8) Wold, S. *Pattern Recognit.* 1976, 8, 127.
- (9) Arcos, J. C. *Environ. Sci. Technol.* 1987, 21, 743.
- (10) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley and Sons: New York, 1978; pp 62-63.
- (11) McGill, J. R.; Kowalski, B. R. *J. Chem. Inf. Comput. Sci.* 1978, 18, 51.
- (12) Wold, S.; Christie, O. *Anal. Chim. Acta* 198, 165, 52.
- (13) van der Voet, H.; Doornbos, D. A. *Anal. Chim. Acta* 1984, 161, 115-123.
- (14) Albano, C.; Dunn, W. J., III; Edlund, U.; Johansson, E.; Norden, B.; Sjöström, Wold, S. *Anal. Chim. Acta* 1978, 103, 429.

Received for review December 9, 1988. Accepted July 31, 1989. Although the research described in this article has been funded by the U.S. Environmental Protection Agency under Cooperative Agreement CR-814363-01-0 with the University of Illinois at Chicago, it has not been subjected to Agency review. It therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. The mention of trade names or commercial products does not constitute endorsement or recommendation for use.